EVALUATING VISUAL CLASSIFICATION MODELS ON OUT-OF-DISTRIBUTION SHIFTS WITH LIMITED TRAINING DATA

A Dissertation
Presented to
The Academic Faculty

By

Aaditya Singh

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in the
College of Computing

Georgia Institute of Technology

May 2023

EVALUATING VISUAL CLASSIFICATION MODELS ON OUT-OF-DISTRIBUTION SHIFTS WITH LIMITED TRAINING DATA

Thesis committee:

Dr. Judy Hoffman, Advisor
School of Interactive Computing
Georgia Institute of Technology

Dr. Zsolt Kira
School of Interactive Computing
Georgia Institute of Technology

Dr. Danfei Xu
School of Interactive Computing
Georgia Institute of Technology

Date approved: April 25, 2023

The sides of the mountain is where things grow, experience is gained and technologies are mastered. The importance of the peak lies only in the fact that it defines the sides.

Dr. APJ Abdul Kalam

ACKNOWLEDGMENTS

I want to thank my advisor, Dr. Judy Hoffman. Thank you for providing the opportunity to hone my interests and skills much further than I imagined before joining Georgia Tech. Through our interactions in project and lab meetings, Judy taught me how to pose research questions and results that can be intriguing and apparent even to an unfamiliar audience. Without her guidance and support, it would be impossible for me and my collaborators to make much of a progress in any of our projects. As a human, I am in awe of her humble and down-to-earth nature and deeply cherish her efforts to make her students happy. I wish her all the accolades and happiness that she richly deserves.

I also want to thank my committee members – Dr. Zsolt Kira and Dr. Danfei Xu. While my interactions with them as a Teaching Assistant might have been brief, it was enough for me to assess that they care deeply about the students and conveying the most precise information. I thank them for putting me in a position where I could stay in touch with the core concepts of the field. I am also grateful to my fellow students with whom I interacted in their courses, I was motivated by their sincere efforts to learn and succeed.

I had the fortune to interact with some highly talented and dedicated peers in and outside Judy's AI Lab. Thanks to Kartik for being an amazing friend and for sticking with our ideas when it mattered the most. Thanks to Prithvi and Viraj for providing insights and guidance whenever required. Thanks to my extremely diligent and intellectual friend Sriram for discussing and collaborating on research ideas. And thanks to my awesome flatmate Krishna, with whom I could share my worries and joys.

My journey in research certainly wouldn't be the same if it weren't for some of my close friends. Thanks to Shreeshail for taking a leap of faith and working with me on our first research projects. As the first steps are often the hardest, much thanks to Shivansh for helping me take those in research and being there for discussing even the tiniest of details. I wish them and all my other friends the best of fortune and success in their endeavors.

Finally, I want to thank my family without whose unwavering support I wouldn't be here in the first place. Thanks to my sister Urvashi and brother-in-law Rajat for being moral supports in a land far away from home. I always had something to look forward to thanks to them and my dear niece Aavya. Thanks to my father Arun Kumar Singh who showed me the value of integrity through his actions, which is crucial for research. Much thanks to my mother Veena Singh, for her indomitable spirit that drove our family forward and put us in a position from where each of us could succeed. Also thanks to my late grandfather Prof. Indrapal Singh, for instilling an early love for Mathematics in me. With their love and support, venturing into the unknown becomes possible for me. I wish them good health and all the joys that the world has to offer.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

As deep learning based models continue to advance several artificial intelligence applications including safety-critical ones, it becomes increasingly important that such models are reliable even under distribution shifts. Moreover, as better models trained on increasingly larger datasets are becoming publicly available, the expectation from a practitioner shouldn't be to train such models on large-scale datasets that might be infeasible to collect.

In this thesis, we focus on two problem settings where out-of-distribution (OOD) performance is measured when the in-domain (ID) training data is limited (i.e. $\sim 10^3 - 10^4$ images) – (1) unsupervised domain adaptation (UDA) where unlabelled OOD data is available for training, and (2) robustness to natural distribution shifts, where OOD data is used for evaluation only. First, we motivate the need for using a more recent family of models (i.e. self-supervised vision transformers) for UDA and briefly describe our method [1] which further improves OOD performance. Second, we describe our recent work [2] on benchmarking robustness to natural shifts with limited ID training data (i.e. low-shot robustness), including the experimental setup and key results.

Overall, the thesis motivates the need for evaluating state-of-the-art deep learning models on diverse out-of-distribution shifts when the amount of training data is limited, by demonstrating that (1) such models can be better utilized for unsupervised domain adaptation and (2) conventional wisdom for out-of-distribution (OOD) robustness (see section 3.3) might not apply when the amount of in-domain training data is not as high.

CHAPTER 1
INTRODUCTION

1.1 Motivation

Deep Learning has taken the world by storm – with a significant rise in the number of applications in the past and this decade, ranging from playing the game of Go [3], self-driving cars [4], to detecting brain tumors [5]. As deep learning based models are increasingly deployed for real-world and often safety critical applications, it is necessary that they perform reliably not only on the dataset(s) used for training and validation, i.e. in-domain (ID) data but also on different kinds of (out-of) distribution shifts that can be expected after deployment. Such models often struggle to generalize to data distributions other than the ones used for training [6, 7, 8]. In this work, we focus on the task of image classification – where the objective is to classify an image into one of many possible classes – and study two kinds of problems related to out-of-distribution (OOD) generalization, i.e. domain adaptation and robustness to natural distribution shifts. We briefly summarize them below.

1.1.1 Domain Adaptation

The goal of domain adaptation (DA) is to transfer a model trained on one (aka source) domain to another (aka target) domain. If the target domain doesn't have class labels for the images, the setting is referred to as unsupervised domain adaptation (UDA) and has been extensively studied [9, 10, 11, 12, 13, 14, 15, 16]. A large body of works leverage models pre-trained on the ImageNet [17] dataset in a supervised manner, i.e. using both the images and the manually annotated class labels. In recent years, however, researchers have shown that models pre-trained on such datasets without using the labels, i.e. via self-supervised learning (SSL) can perform on-par or better than the supervised ones on downstream tasks [18, 19, 20, 21, 22]. Note that most recent SSL methods [21, 22, 23] use Vision Trans-

formers (ViTs), that are more computationally efficient [24] and also have better OOD calibration than Convolutional Neural Networks (CNNs) [25]. While CNNs have largely been used for UDA and some works have studied self-supervised adaptation [26, 27], no work thus far has focused on using SSL ViTs for UDA and whether their performance can be further improved. Thus, we aim to answer the following questions in our work [1]:

Q1. *Do recent UDA methods also improve performance with self-supervised (SSL) ViTs?*

Q2. *Can the emergent properties of SSL ViTs [21] lead to a better adaptation method?*

1.1.2 Robustness to Natural Distribution Shifts

Depending on the domains, assuming access to additional data from the target domain for subsequent model training might not be a fair assumption. For instance, in the case of detecting rare animal species [28] or the presence of tumors in different hospital scans [29], it's unreasonable to expect the practitioner to collect substantial amount of target data after model deployment. Therefore, robustness studies impose a harsher constraint on the out-of-distribution (OOD) generalization problem by *not* assuming access to the target OOD data, and using it only for evaluation purposes.

Previous works have studied such OOD generalization capabilities of models under synthetic [30, 31, 32, 33, 34, 35] and natural distribution shifts [36, 37, 38, 39, 40, 41]. Notably, [42] find that robustness interventions used for synthetic shifts offer little to no improvements for natural shifts through a large-scale study of several supervised models. Recent methods that provide robustness improvements to such natural shifts utilize self-supervised [22, 23] or large-scale vision-language pre-trained models such as CLIP [43] and perform fine-tuning with fully labelled ID datasets [22, 23, 44, 45, 46]. Crucially, such fine-tuning can be resource intensive and access to large and labelled datasets can be infeasible, as discussed previously. Thus, in our work [2] we perform a study of robustness to natural shifts in the low-shot regimes – spanning datasets, architectures, pre-trained initializations, and state-of-the-art interventions – and aim to answer these questions:

Q3. *Does a single model provide better robustness across datasets in low-shot regimes?*

Q4. *Does robustness in the full-shot regime imply that in the low-shot regimes?*

1.2 Thesis outline

The rest of the thesis is aimed towards answering the questions raised in subsection 1.1.1 and subsection 1.1.2, and is structured as follows.

- In Chapter 2, we focus on domain adaptation with self-supervised ViTs. We describe the baselines and some key results from our work [1] and provide additional insights.
- In Chapter 3, we explain our recent work [2] on low-shot robustness to natural shifts. We briefly describe the experimental setting, evaluation metrics, and key results.
- In Chapter 4, we discuss some of the key issues and limitations that we observe in different works along with potential future directions to alleviate them.
- In Chapter 5, we conclude our findings to motivate better analysis and evaluation of visual classification models on out-of-distribution shifts with limited training data.

CHAPTER 2

ADAPTING SELF-SUPERVISED VISION TRANSFORMERS

Recall from Chapter 1 that methods for unsupervised domain adaptation (UDA) typically leverage supervised Convolutional Neural Networks (CNNs) or Vision Transformers (ViTs), and it remains unknown whether such UDA methods provide improvements for self-supervised (SSL) ViTs. We discuss these methods and models in the following section.

2.1 Related Work

Notable paradigms for UDA include (1) domain adversarial learning [11] which aims to learn a feature representation space that is domain invariant and class discriminative, and (2) selective self-training [47] which uses a model trained on source domain to obtain pseudo labels for the target domain, and selectively increase model's confidence based on some criterion. While our work [1] follows the latter paradigm, we adopt both kinds of methods for comparison and summarize them below.

- **CDAN [13].** CDAN captures cross-covariance between feature embeddings and classifier predictions for better class discriminativeness in domain adversarial learning.
- **MCC [48].** Minimum Classifier Confusion (MCC) uses the model predictions on target domain to minimize pair-wise class confusion for aligning domains in a non-adversarial fashion.
- **SENTRY [1].** SENTRY is a selective self-training approach which increases model confidence on "reliable" target instances and decreases it on "unreliable" ones. The reliable target instances are the ones for which the model prediction is consistent across randomly augmented versions of that instance.

- **Shen *et. al* [27].** They perform contrastive learning [18, 19] on the pooled source and target domains followed by fine-tuning on the source domain. We adopt their approach by performing method-specific SSL followed by source fine-tuning.

- **TVT [16].** TVT is a UDA method for supervised ViTs which injects a module for learning transferability into ViT's attention blocks. It also performs global domain adversarial alignment and class discriminative clustering.

Compared to these approaches, our method (see section 2.2) is explicitly meant for adapting some of the most recent self-supervised ViTs such as MAE [22] and DINO [21]. In general, self-supervised learning (SSL) is performed by designing a proxy task for training models, with the goal of learning feature representations that are useful for many downstream tasks. We briefly summarize the SSL approaches for MAE and DINO below.

- **MAE [22].** MAE uses the proxy task of masked auto-encoding, i.e. predicting missing image patches given the remaining ones. The authors find that a large masking ratio of $\sim 75\%$ not only leads to better feature representations but also meaningful reconstruction of the missing image patches.

- **DINO [21].** DINO jointly trains a student model which sees local and global augmented views of an image, to match the predictions of a teacher model which only sees the global augmented views. The authors find that aside from useful feature embeddings, other properties such as object localization also emerge from this process.

2.2 Method

We briefly describe our method for adapting SSL ViTs which we call Probing Attention-Conditioned Masking Consistency or `PACMAC` in this section. The method has 3 stages: (1) With ImageNet pre-trained SSL ViTs, an additional pre-training step using the same SSL strategy is performed over the pooled source and target domains. (2) The pre-trained model

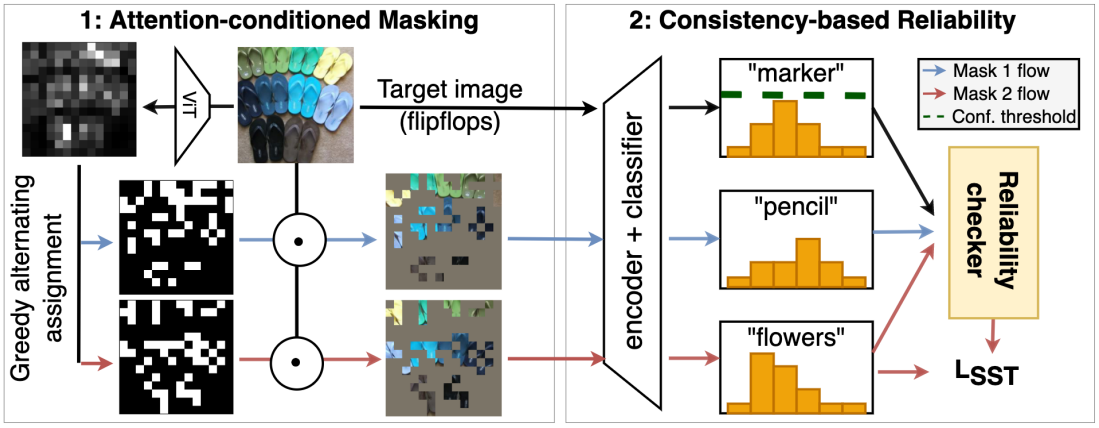

Figure 2.1: **Overview of** PACMAC. **Left.** Model's attention on the target image is used to generate disjoint masks that retain highly attended patches of the input image via greedy allocation strategy. **Right.** Next, the model's predictive consistency between original and masked images is employed to select target instances for self-training.

is fine-tuned in a supervised fashion with the labelled source images. (3) Finally, a selective self-training step is performed in which model's confidence is increased on "reliable" target instances. Reliability is determined by checking whether (a) model's confidence for an augmented view of an image is above a certain threshold T, and (b) model's predictions are consistent for that view and k randomly augmented and masked (with ratio mr) views. The masks are obtained by leveraging (SSL) ViT's attention mechanism, wherein the image patches are sorted in descending order of attention weights and allocated to the masks in a greedy round-robin fashion. This strategy allows each mask to retain some of the highly attended patches, and we leverage the observations from prior work [21] and our experiments that such patches in SSL ViTs are often semantically meaningful. These masks are then applied to the k augmented views of the target image, as shown in Figure 2.1.

We find that PACMAC performs better or on-par with the competing methods on average across shifts in the OfficeHome [49], DomainNet [50, 51], and VisDA [52] benchmarks across initializations. Results with MAE and DINO on OfficeHome are shown in Table 2.1.

Table 2.1: **Target accuracies.** Results with MAE [22] and DINO [21] on OfficeHome [49] are shown. PACMAC performs better or on-par with the competing methods.

IN1K Init.	Method	A → C	A → P	A → R	C → A	C → P	C → R	P → A	P → C	P → R	R → A	R → C	R → P	AVG
MAE [22]	source	46.4	57.6	71.0	51.1	60.0	62.6	51.4	46.9	70.5	66.3	52.2	77.2	59.4
	CDAN [13]	45.3	58.8	69.1	51.6	60.7	61.5	53.4	45.5	72.4	67.7	49.9	78.0	59.5
	MCC [48]	43.9	61.2	70.7	52.8	59.9	62.8	51.1	40.3	70.9	66.2	48.3	76.3	58.7
	Shen *et al.** [27]	57.1	63.6	71.9	57.9	65.6	67.1	55.5	56.7	71.2	69.0	62.6	79.4	64.8
	SENTRY [47]	54.8	65.6	74.4	56.5	65.8	69.8	57.6	54.9	75.5	68.9	60.0	81.6	65.5
	PACMAC	58.9	68.2	74.1	60.6	67.1	67.2	57.3	59.2	74.4	68.6	63.9	82.7	**66.8**
DINO [21]	source	53.1	65.0	75.2	62.0	66.2	70.4	60.8	50.5	77.0	72.8	53.9	81.2	65.7
	CDAN [13]	49.0	70.0	76.4	60.0	67.3	71.2	64.7	47.0	79.9	75.1	56.4	81.8	66.5
	MCC [48]	44.4	74.2	79.6	61.9	67.6	72.4	63.0	40.1	79.2	73.3	47.1	82.8	65.5
	TVT [16]	48.3	65.7	73.6	60.6	68.8	64.6	57.1	44.1	75.4	71.0	53.7	77.2	63.3
	Shen *et al.** [27]	53.1	69.4	76.7	62.6	68.9	71.4	62.2	51.8	76.0	73.5	56.3	81.8	67.0
	SENTRY [47]	59.5	72.0	76.8	66.1	71.1	73.4	63.7	56.2	77.8	72.4	63.0	81.9	69.5
	PACMAC	54.9	74.7	79.3	65.7	74.0	74.5	63.3	55.8	79.2	73.1	58.4	83.9	**69.7**

2.3 Other Experiments: Domain Translation

We admit that our development of an UDA method for SSL ViTs didn't start with a selective self-training approach. Rather, we focused on the image reconstruction properties of MAE [22] from partial inputs. The core idea can be summarized as follows: (1) Train domain specific MAE decoders to perform reconstruction separately for source and target images. The expectation is for the decoders to capture domain specific *style* information. (2) Instead of optimizing cross-entropy loss for the source image, *translate* the source image into the style of the target image with the help of target decoder. The translation step could be performed with the help of works in neural style transfer (NST) literature.

Note that a large number of NST works use CNNs (notably VGG-19 [53]) as the de-facto model family [54, 55, 56, 57, 58]. The choice of CNNs is informed by their increasingly complex and hierarchical feature representations [59, 60], which is different from how ViTs transfer information across the layers [24, 61]. To the best of our knowledge, works that use ViTs for NST [62] still rely on the same VGG based perceptual losses that involve tuning many hyperparameters on much larger datasets [63, 64]. Concurrent work which uses domain specific decoders for multi-source adaptation [65] also relies on several other components, and their reconstructions still don't seem to capture domain specific in-

Figure 2.2: **Stylization Results.** We show some stylized images obtained via WCT [56] with VGG-19 [53] and ViTB-16 [66]. WCT captures the general color palette and textures well with VGG-19 (see footnote 2) but not with ViTB-16.

formation. Thus, while the idea discussed in this section might be interesting, it isn't well aligned with the findings in NST literature and remains hard to implement.

Nonetheless, we perform an oracle experiment with a supervised ViTB-16 [66] by following the general approach of WCT [56]. We briefly summarize this approach as follows: (1) Freeze the pre-trained ViT encoder and train a ViT decoder [22] to perform image reconstruction [1]. (2) Extract the feature representations from n^{th} block of the encoder for the two source and target images. (3) Apply a feature transform such as AdaIN [55] or WCT [56] to transfer the "style" of target image onto source image in the feature space. (4) Reconstruct the stylized image from the stylized feature representation by passing it to the pre-trained decoder. We use $n = 1$ as it led to more visually appealing results and show some stylization results in Figure 2.2. While WCT captures the general color palette and textures from the target images well with VGG-19 [2], it largely fails to do so for ViTB-16. We omit the results with AdaIN [55] as they often resemble the source image itself.

[1] We pre-train the IN1k initialized MAE decoder [22] on the Clipart and Product domains in Office-Home [49] dataset, but find that the reconstructed images for other domains also match the input images.

[2] WCT [56] is applied at several layers of VGG-19 for enhanced stylization, whereas we only apply it to the feature representations obtained from the last block of ViTB-16. However, even last layer stylization results as shown in [56] are much more visually appealing than the ones obtained for ViTB-16 in Figure 2.2.

Overall, we focus on leveraging Vision Transformers (ViTs) pre-trained with state-of-the-art self-supervision (SSL) for the unsupervised domain adaptation (UDA) task. We implement and benchmark existing UDA methods for SSL ViTs, and propose a selective self-training based method called `PACMAC` which relies on predictive consistency across attention-seeded masked views of target images. Our results demonstrate that `PACMAC` performs better or on-par with competing methods on standard benchmarks.

CHAPTER 3

MEASURING LOW-SHOT ROBUSTNESS TO NATURAL SHIFTS

Recall from Chapter 1 that out-of-distribution (OOD) robustness studies and methods leverage large amounts of in-domain (ID) labelled data for model fine-tuning, which can be prohibitive for practitioners due to resource constraints and the nature of datasets such as rare animal species [28] and hospital scans [29]. Thus, in our work [2] we formulate the "Low-Shot Robustness" setting in which: (1) We assume access to a model pre-trained on large scale datasets such as ImageNet [17] (IN1k), and fine-tune the pre-trained model along with a classifier head with the limited ($\sim 10^3 - 10^4$ images) ID training data. (2) The fine-tuned model is then evaluated on the OOD test data. The setting is also visually depicted in Figure 3.1. We briefly summarize the datasets and data regimes, pre-trained models, and fine-tuning methods that we include in our testbed in the following section.

3.1 Experimental Setting

3.1.1 Datasets & Data Regimes

ImageNet [17] (IN1k). Previous works [22, 23, 67] often measure OOD robustness by training on full IN1k and testing on some or all of these 5 distribution shifts, i.e. IN-R [40], IN-S [38], IN-A [39], INv2 [36], and ObjectNet [37]. We include all the 5 distribution shifts and the low-shot subsets with $1, 5,$ and 10 images per class subsets provided by [23]. For validation, we use the IN1k val split and report top-1 accuracy. For testing, we follow previous works [45, 46] and report top-1 accuracy averaged on the 5 distribution shifts.

iWildCam [28]. The iWildCam dataset consists of images of 182 animal species captured by different camera traps that are considered distribution shifts. We use the WILDS benchmark [41] to create low-shot subsets with images in $1\%, 10\%,$ and 20% ratio from

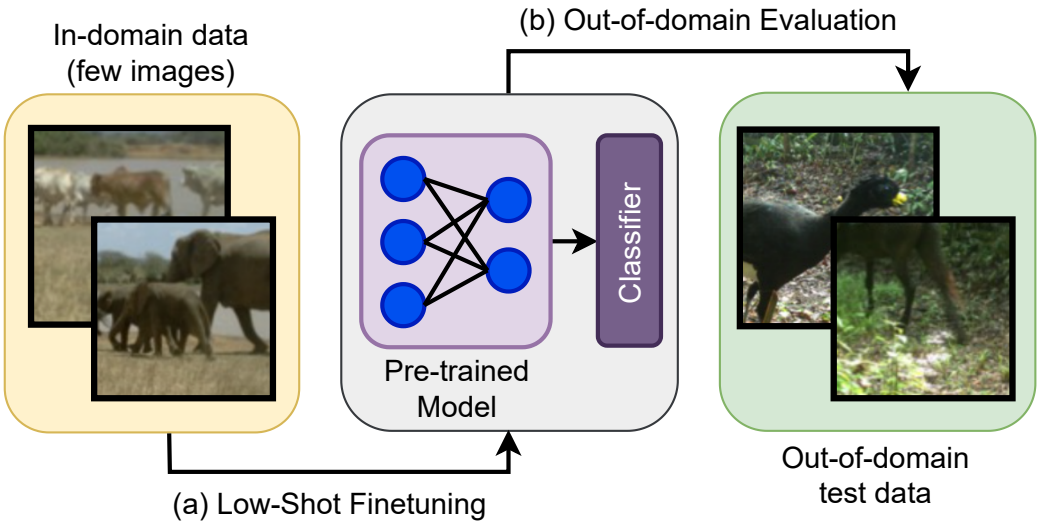

Figure 3.1: **Low-Shot Robustness Setting.** (a) We assume access to a model pre-trained on large scale datasets (e.g. IN1k), attach a classifier head on top and fine-tune the model with the few labelled in-domain (ID) images. Different methods for fine-tuning are used that demonstrate robustness when there is typically order of magnitudes higher training data. (b) The (low-shot) fine-tuned model is then evaluated on out-of-domain (OOD) data.

`train` shift for training, given the imbalanced class distribution in the dataset. For validation, we use the `val-id` shift which has 7314 images. For testing, we the `val-ood` shift which has $14,961$ images. We report per-class accuracy for validation and testing.

Camelyon [28]. The Camelyon dataset consists of histopathological scans that may or may not contain tumor tissue, i.e. 2 classes. The scans are obtained from different hospitals that are considered distribution shifts. We again use the WILDS benchmark [41] to create low-shot subsets with $1500, 7500,$ and 15000 images per class `train` shift for training, as the shifts are well balanced. For validation, we use the `val-id` shift which has $33,560$ images. For testing, we the `val-ood` shift which has $34,904$ images. We report per-class accuracy for validation and testing.

We show sample images from each dataset in Figure 3.2 and refer to the 3 low-shot regimes discussed previously as *extreme*, *moderate*, and *high* low-shot regimes respectively.

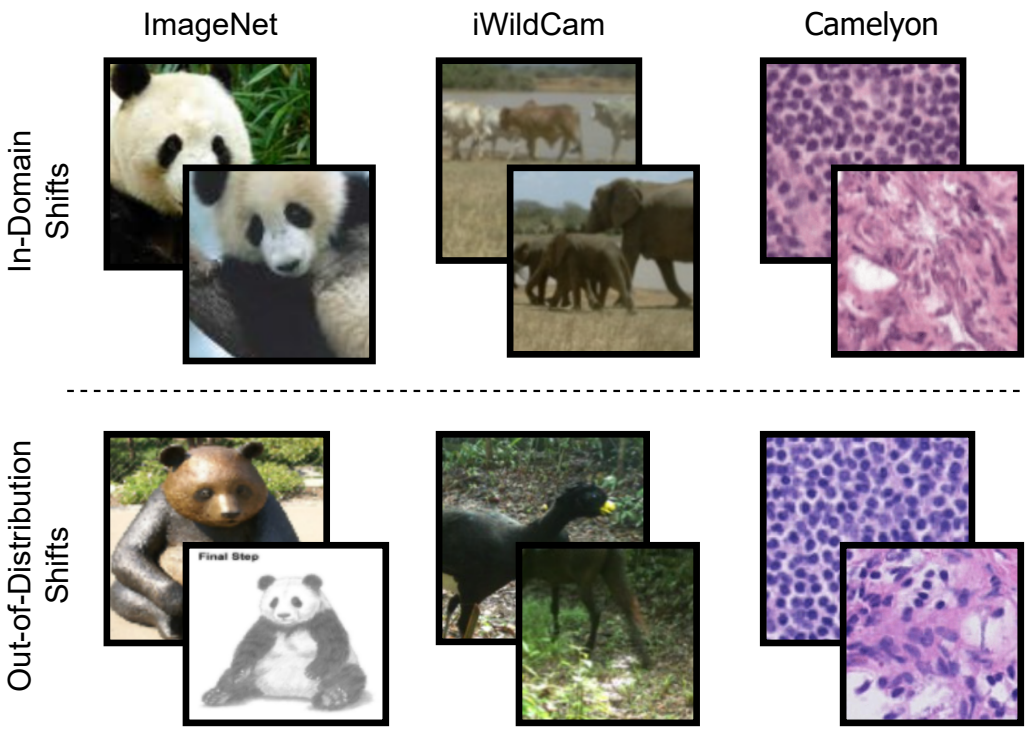

Figure 3.2: **Datasets & Distribution Shifts.** Sample images from ImageNet [17] and some of the associated distribution shifts [40, 38], iWildCam [28], and Camelyon [29] datasets.

3.1.2 Pre-trained Models

We refer to ImageNet pre-trained models as *standard* models, i.e. trained without additional interventions or larger datasets. We include the following self-supervised CNNs – SwAV [68], DINO [21] and ViTs – DINO [21] and MSN [23] as standard models. For datasets other than ImageNet, we also include the following supervised CNNs – RN50 [69] and ViTs – DEIT [66]. Note that ImageNet supervised models violate the "low-shot" condition on ImageNet as they have already been trained with all of the labels. We mostly use the ViTS-16, ViTB-16, RN50, and RN50w2 model sizes for the different datasets.

The BS-CDFSL [70] study shows that simpler transfer learning baselines outperform meta-learning approaches on the cross-domain few-shot learning task, hence we compare the 3 classifiers – Logistic Regression [71, 23], Mean-Centroid Classifier [72], and Baseline++ [73] – and choose the best performing ones. While Logistic Regression performs well on ImageNet and iWildCam datasets, Baseline++ performs better on Camelyon.

3.1.3 Fine-tuning Methods

Recent works that achieve impressive robustness gains on some of the datasets and distribution shifts (see subsection 3.1.1) in the full-shot regime either perform (1) strategic fine-tuning (LP-FT [44], RobustViT [67]) or (2) weight-space ensembling (WiSE-FT [45], Model Soups [46]). We incorporate all of these methods in our testbed as robustness interventions and briefly describe them below.

- LP-FT [44]: LP-FT follows a two-stage training process in which a randomly initialized linear head is trained first followed by fine-tuning the entire model (with the trained head) end-to-end.

- WiSE-FT [45]: WiSE-FT uses a weight-space ensemble to combine a zero-shot model such as CLIP [43] with the fully fine-tuned model. For IN1k pre-trained models, we ensemble LP and LP-FT checkpoints due to absence of a zero-shot head.

- Model Soups [46]: Model Soups uses a weight-space ensemble of a linear-probed model that is then trained with randomly sampled epochs, learning rates, weight decay, label smoothing [74], mixup [75], and RandAugment [76].

- RobustViT [67]: RobustViT employs an unsupervised object localization method such as TokenCut [77] to generate offline segmentation maps. A supervised ViT is then trained such that it's saliency maps [78] resemble the offline ones and its classification accuracy is maintained.

Note that most of these methods use vision-language models such as CLIP [43], whereas we adopt them for both CLIP and ImageNet pre-trained initializations. We also include zero-shot CLIP [43, 45] as a robustness intervention owing to its strong performance on robustness benchmarks. For RobustViT, we first perform a linear-probing step for self-supervised ViTs on ImageNet, but the method remains hard to implement for other and especially non object-centric datasets due to its requirement of offline segmentation maps.

3.2 Evaluation Metrics

Evaluating the absolute performance of a model using out-of-distribution (OOD) shifts may indicate robustness, but does not take into account the model's in-domain (ID) performance. As noted by [42], models with similar OOD performance may have significantly different ID performances. A more comprehensive definition of robustness should consider OOD performance beyond what is expected from achieving a certain level of ID performance. Therefore in addition to comparing absolute performance to measure robustness, we adopt the *effective* and *relative* robustness framework used by prior works [36, 42, 45]. We briefly describe how the associated metrics are computed below.

For measuring effective robustness, a baseline OOD accuracy for a given ID accuracy x is obtained by fitting a log-linear curve $\beta(x)$ over the set of ID and OOD accuracies of standard models, i.e. $\{f_1^s, f_2^s, \dots f_n^s\}$ where $f^s = (acc_{id}^s, acc_{ood}^s)$. The curve is defined as:

$$\beta(x) = \sigma(w \text{ logit}(x) + b) \tag{3.1}$$

where $\text{logit}(x) = \ln \frac{1}{1-x}$ and σ is its inverse. In practice, $\beta(x)$ is obtained by transforming each point $(x, y) \rightarrow (\text{logit}(x), \text{logit}(y))$ and solving linear regression. To visualize, (acc_{id}^s, acc_{ood}^s) are plotted on a scatter plot where x and y axes denote the ID and OOD accuracies respectively.

Once $\beta(x)$ is obtained, effective robustness of an *intervention*[1] r applied on the model f^s, i.e. $f^r = (acc_{id}^r, acc_{ood}^r)$ is defined as:

$$\rho(f^r) = acc_{ood}^r - \beta(acc_{id}^r) \tag{3.2}$$

which describes whether the intervention leads to an OOD accuracy beyond what can be

[1] We note that for models pre-trained on large external datasets such as CLIP [43], it is questionable what kind of datasets constitute in or out-of-distribution. Thus, we treat it as an intervention that isn't included in the standard set of models (see subsection 3.1.2) used to compute effective and relative robustness.

expected from achieving a higher ID accuracy.

As noted by [42], an intervention can result in a high positive $\rho(f^r)$, indicating effective robustness, but it can still decrease both ID and OOD accuracies which is obviously not desirable. Thus, along with effective robustness, we also measure relative robustness which is defined as:

$$\tau(f^r) = acc^r_{ood} - acc^s_{ood} \tag{3.3}$$

Following [42], an intervention is said to improve robustness if it is both effectively and relatively robust, i.e. $\rho(f^r) > 0$ and $\tau(f^r) > 0$. As seen in our experiments, interventions often lack simultaneous effective and relative robustness across different low-shot regimes. We refer to $\rho(f^r)$ as ρ and $\tau(f^r)$ as τ for simplicity.

3.3 Results

From existing literature on self-supervised learning (SSL) and out-of-distribution (OOD) robustness, we seem to arrive at the following conclusions for robustness in the full-shot regime: (1) Amongst IN1k pre-trained models, SSL ViTs are more robust with the recent ones being better [22, 23]. (2) Without additional interventions, zero-shot models such as CLIP [43] provide superior robustness than ImageNet pre-trained ones. (3) The robustness of such models can be improved further with recent robustness interventions [44, 45, 46]. In our work [2], we question each of these observations for robustness in the low-shot regimes described in subsection 3.1.1. We briefly summarize our findings below. [2]

3.3.1 ImageNet Pre-Trained Model Comparison

We compare ImageNet (IN1k) pre-trained models with similar number of trainable parameters in low-shot regimes with respect to absolute performance on in-domain (ID) and

[2]We train the models to near completion, i.e. $98 - 100\%$ training accuracy and select the checkpoint with the best in-domain (ID) validation performance. We follow prior works [23, 41, 45] for design choices and hyperparameters, and perform a grid search over epochs, learning rates, and weight decay whenever feasible.

Table 3.1: **Comparison of IN1k pre-trained self-supervised (SSL) ViTs.** No single initialization or model size outperforms others on average across low-shot regimes on the different datasets.

	ImageNet		iWildCam		Camelyon	
	ID	OOD	ID	OOD	ID	OOD
1 MSN ViTS-16 [23]	58.99	21.51	26.41	**19.99**	83.62	75.67
2 DINO ViTS-16 [21]	53.78	19.09	24.78	19.75	**88.08**	**85.09**
3 MSN ViTB-16 [23]	**61.40**	**22.81**	24.78	19.65	86.40	78.84
4 DINO ViTB-16 [21]	56.72	21.98	**27.40**	19.82	86.93	84.33

out-of-distribution (OOD) shifts in Figure 3.3. We see that self-supervised (SSL) ViTs generally outperform CNNs and the supervised counterparts on both ID and OOD shifts.

Next, we vary the initialization and model size of IN1k pre-trained SSL ViTs and show the average performance across low-shot regimes in Table 3.1. No single initialization or model size works better for the different datasets, with MSN ViTB-16 performing better on both ID and OOD shifts on ImageNet and DINO ViTS-16 on Camelyon. Thus, while SSL ViTs are more robust than CNNs and supervised counterparts, no single initialization or model size works better in low-shot regimes across datasets.

3.3.2 Pre-training Data Scale and Strategy

In Table 3.2, we compare IN1k pre-trained SSL ViTs with the models pre-trained on larger datasets where applicable, i.e. CLIP ViT [43] and ImageNet-21k [79] (IN21k) supervised ViT [24]. For a fair comparison, we use the ViTB-16 architecture for all models. It can be seen that while zero-shot CLIP performs significantly better on ImageNet, IN1k pre-trained DINO outperforms other models on iWildCam and Camelyon. Thus, without additional robustness interventions and in low-shot regimes, models pre-trained on large and external datasets such as CLIP [43] provide superior robustness than ImageNet pre-trained models on ImageNet, but not on other datasets such as iWildCam and Camelyon.

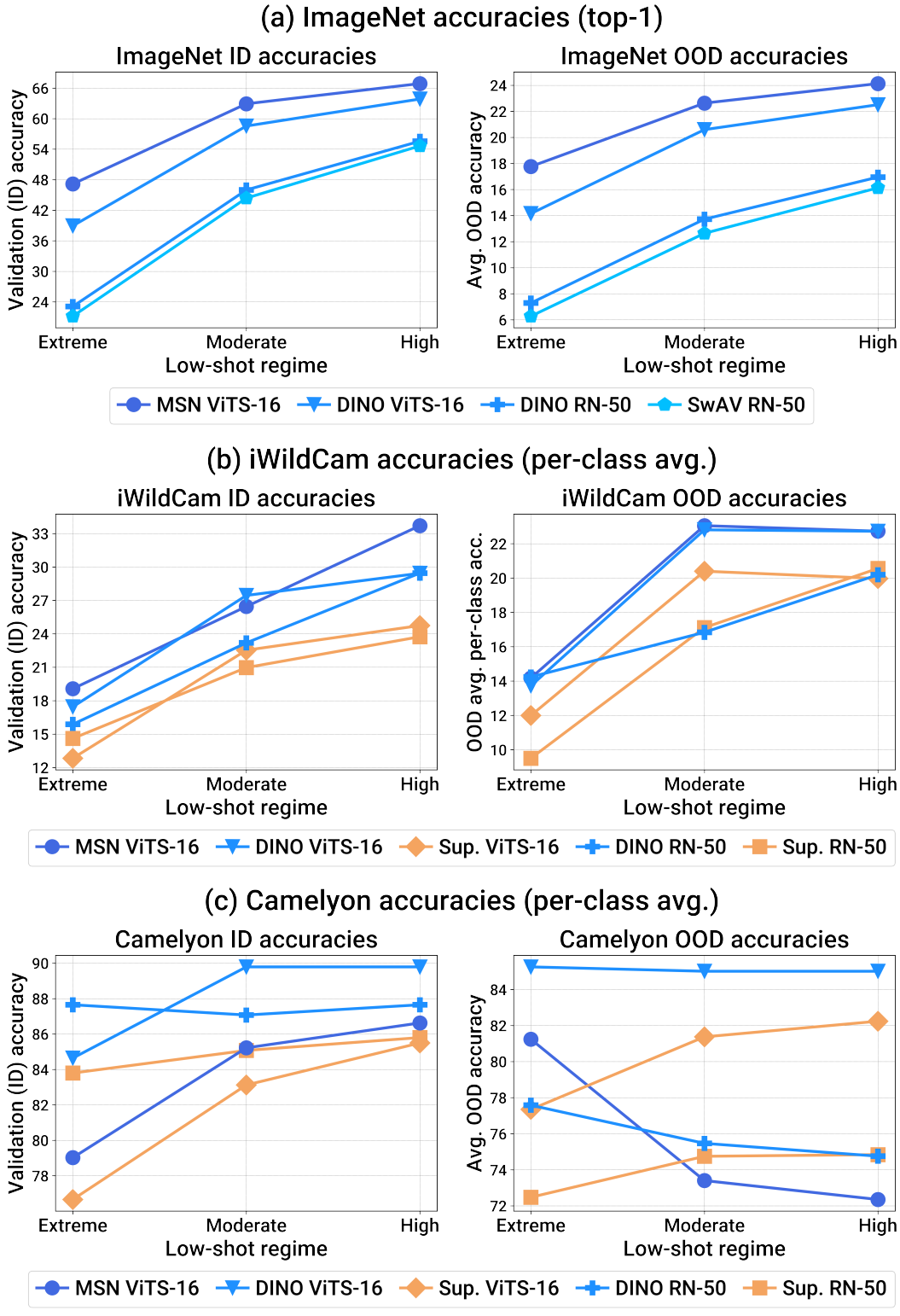

Figure 3.3: **Comparison of IN1k pre-trained architectures and initializations.** With similar number of parameters, self-supervised (SSL) ViTs generally perform better on both ID and OOD shifts compared to SSL CNNs and the supervised counterparts where applicable.

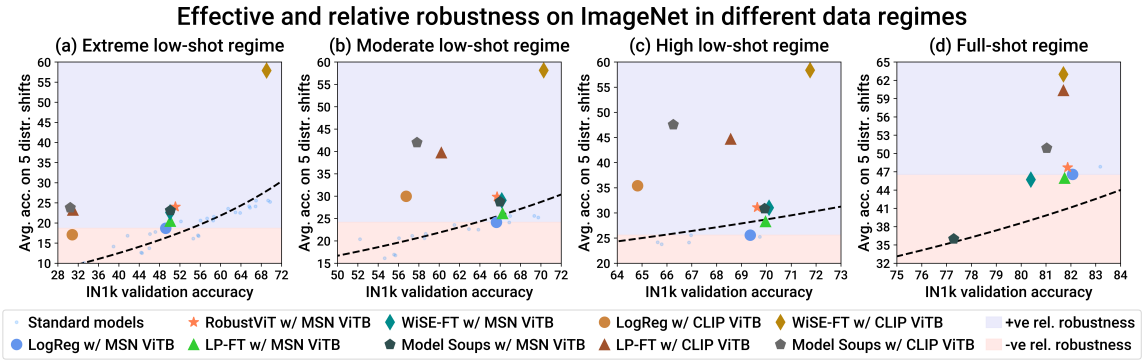

Figure 3.4: **Effect of robustness interventions on ImageNet.** Plots (a), (b), and (c) show performance of interventions in low-shot regimes (subsection 3.1.1). Plot (d) shows performance of interventions in the full-shot regime. Interventions located above the line ($\rho > 0$) and in the blue region ($\tau > 0$) are said to improve robustness (section 3.2). Interventions largely improve robustness in low-shot regimes with MSN ViTB-16 and in all data regimes with CLIP ViTB-16.

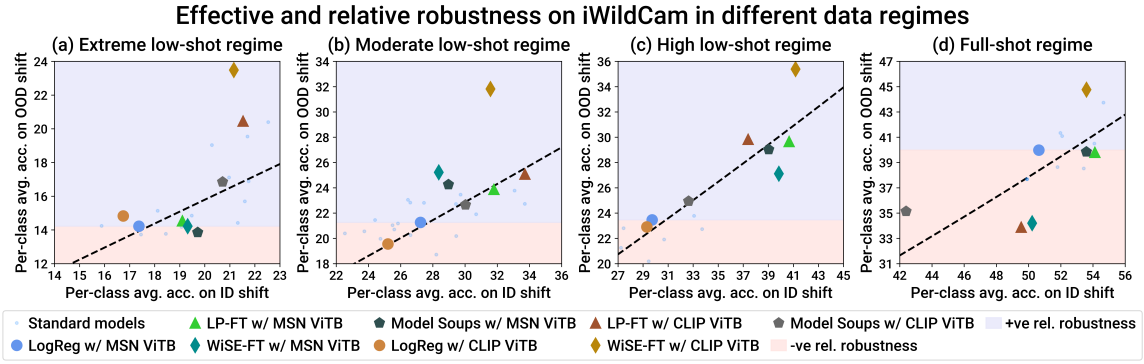

Figure 3.5: **Effect of robustness interventions on iWildCam.** Interventions often fail to improve robustness in both the full and low-shot regimes with MSN ViTB-16. Only WiSE-FT with CLIP ViTB-16 improves robustness in all data regimes.

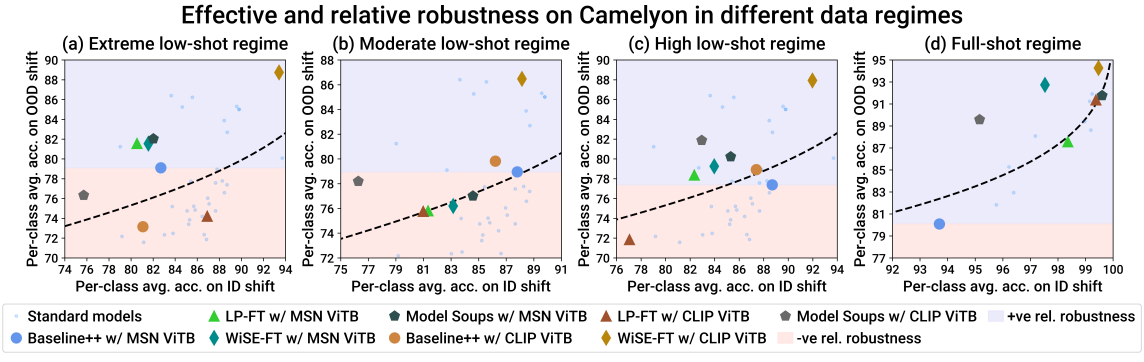

Figure 3.6: **Effect of robustness interventions on Camelyon.** Interventions often improve robustness in the full-shot regime with both MSN and CLIP ViTB-16 but fail to do so in *extreme* or *moderate* low-shot regimes, except WiSE-FT with CLIP ViTB-16.

Table 3.2: **Comparison between ViTs pre-trained on different datasets.** On average across low-shot regimes, ImageNet pre-trained SSL ViT's such as DINO are worse than CLIP on both ID and OOD shifts on ImageNet. However, it performs significantly better than CLIP and ImageNet-21k supervised ViT on iWildCam and Camelyon datasets.

	ImageNet		iWildCam		Camelyon	
	ID	OOD	ID	OOD	ID	OOD
1 CLIP zero shot [43, 45]	**67.93**	**57.37**	9.67	16.82	50.48	51.55
2 CLIP [43]	50.8	27.50	23.75	19.10	84.9	77.3
3 Supervised (IN21k) [24]	N/A	N/A	16.84	16.90	85.18	81.07
4 Supervised (IN1k) [66]	N/A	N/A	22.27	18.57	83.35	83.24
5 MSN (IN1k) [23]	61.40	22.81	24.78	19.65	86.40	78.84
6 DINO (IN1k) [21]	56.72	21.98	**27.40**	**19.82**	**86.93**	**84.33**

3.3.3 Effect of Robustness Interventions

We now use the effective and relative robustness framework (see section 3.2) for observing the effect of robustness interventions described in subsection 3.1.3. Unless stated otherwise, we use MSN as a reference and ViTB-16 architecture for a fair comparison. We apply these interventions on both MSN and CLIP and present the dataset-wise observations for ImageNet in Figure 3.4, iWildCam in Figure 3.5, and Camelyon in Figure 3.6.

Most interventions provide large robustness improvements on ImageNet with both MSN and CLIP in all data regimes. Whereas on iWildCam, interventions are often not effectively robust with MSN and only WiSE-FT [45] with CLIP improves robustness in both the full and low-shot regimes. On Camelyon, most interventions improve relative robustness in the full-shot regime with MSN and both effective and relative robustness with CLIP. However, except WiSE-FT with CLIP, they fail to do so in the *extreme* or *moderate* low-shot regimes. Thus, with additional interventions, robustness in the full-shot regime doesn't imply that in the low-shot regimes across different datasets such as iWildCam and Camelyon.

Overall, we study robustness to several natural distribution shifts in low-shot regimes, which addresses the gap in the literature and marks the first in-depth study of its kind. From our evaluations, we observe that: (1) Amongst ImageNet pre-trained initializations, self-supervised ViTs are often more robust in low-shot regimes across different datasets but the

best initialization or model size is dataset dependent. (2) Without additional interventions, models pre-trained on large external datasets such as CLIP can be much more robust on ImageNet but not on other datasets such as iWildCam and Camelyon. (3) Depending on the initialization, robustness interventions fail to improve robustness in the full-shot regime or in different low-shot regimes on such datasets. These results demonstrates that conventional wisdom for robustness to natural distribution shifts in the full-shot regime (see section 3.3) might not apply in the low-shot regimes. We hope to motivate researchers to focus on this problem of practical importance.

CHAPTER 4
LIMITATIONS AND DISCUSSION

We note that there are some key issues and limitations in different works on unsupervised domain adaptation (UDA) and our work on low-shot robustness to natural shifts. We discuss them in the following sections, and later present our thoughts on the broader out-of-distribution (OOD) generalization problem informed by these works.

4.1 PACMAC and unsupervised domain adaptation

While PACMAC [1] works better than competing methods with self-supervised (SSL) ViTs such as MAE [22] and DINO [21], we find that it heavily underperforms with supervised ViT compared to methods such as TVT [16]. Interestingly however, the label overlap between ImageNet-21k [79] and the standard UDA benchmarks, i.e. DomainNet [50, 51], OfficeHome [49], and VisDA [52] is almost 100%. This is concerning because the UDA setting assumes that labelled images from target domain are inaccessible, which doesn't seem to hold in practice. Thus, SSL models could be a fairer initialization choice.

Researchers have also shown that the performance-wise order of UDA methods is heavily dependent on pre-trained initializations [15] and datasets [41]. Also, it is possible that methods such as SENTRY [47] can perform better with in-domain SSL pre-training and a more extensive search over hyperparameters. Therefore, we believe that our use of SSL ViTs and benchmarking existing UDA methods with them is an equally (if not more) important contribution than PACMAC itself.

4.2 Low-shot robustness to natural distribution shifts

Several other directions could be explored for robustness to natural distribution shifts in low-shot regimes. First, collecting unlabelled in-domain data might be feasible even for settings such as iWildCam [28] and Camelyon [29]. As seen from recent works in UDA [27, 1], in-domain SSL pre-training can provide significant improvements with unlabelled OOD data. However, pre-training schedules can be time and resource consuming [20] and the objectives of SSL methods might not be suitable for class imbalanced datasets [80].

Second, out-of-distribution (OOD) performance might be sensitive to different kinds of augmentations and loss functions. While we implicitly incorporate some of them as a part of robustness interventions such as Model Soups [46], separately analyzing their effect on different datasets and data regimes could be an interesting direction for future work.

4.3 Thoughts on out-of-distribution generalization

Computer vision community has studied out-of-distribution (OOD) generalization (see [81]) in related but distinct problem settings, such as unsupervised domain adaptation and robustness to distribution shifts. However, differences in factors such as datasets and pre-trained models make it difficult to arrive at any general conclusions for OOD generalization. Development and adoption of benchmarks such as WILDS [41] that control for such factors can greatly improve our understanding and lead to better generalization methods.

Nonetheless, vision-language models pre-trained on large-scale datasets such as CLIP [43] provide unprecedented robustness gains to several OOD shifts, especially when combined with recent fine-tuning methods [44, 45, 46] that can possibly be augmented with access to additional unlabelled data. Recent work [82] also attempts to make the zero-shot decisions of such models more interpretable with the help of prompts obtained from large language models such as GPT-3 [83]. Combining such techniques with the advanced fine-tuning methods could be a highly interesting and useful direction for future work.

CHAPTER 5
CONCLUSION

We now conclude this thesis on evaluating visual classification models on out-of-distribution (OOD) shifts with limited training data. Through our experiments described in Chapters 2 and 3, we provide short answers to the key questions raised in Chapter 1, section 1.1.

Q1. *Do recent UDA methods also improve performance with self-supervised (SSL) ViTs?*
A: While some UDA methods fail to improve upon the source-only baseline with self-supervised (SSL) ViTs [21, 22], our implementations of recent methods [27, 47] that were primarily meant for CNNs also provide improvements for SSL ViTs.

Q2. *Can the emergent properties of SSL ViTs [21] lead to a better adaptation method?*
A: Yes. We propose a novel selective self-training approach called `PACMAC` in our work [1] which uses the SSL ViT's attention mechanism for masking based consistency. `PACMAC` performs better or on-par with the competing methods on standard benchmarks.

Q3. *Does a single model provide better robustness across datasets in low-shot regimes?*
A: No. Amongst ImageNet (IN1k) pre-trained models, SSL ViTs [21, 23] are generally more robust than CNNs [68, 21] and supervised counterparts [69, 66], but no single model performs better across datasets. Similarly, CLIP [43] performs significantly better than other models on both ID and OOD shifts on ImageNet, but IN1k pre-trained DINO [21] outperforms others on iWildCam [28] and Camelyon [29] datasets.

Q4. *Does robustness in the full-shot regime imply that in the low-shot regimes?*
A: No. While most robustness interventions [44, 46, 67] largely improve robustness in both the full and low-shot regimes on ImageNet, depending on the initialization they fail to do so in the full-shot or in different low-shot regimes on iWildCam and Camelyon.

We firmly believe that state-of-the-art deep learning models should be evaluated on diverse out-of-distribution (OOD) shifts when the amount of data available for fine-tuning is limited (i.e. $\sim 10^3 - 10^4$ images) as it reflects a practical scenario which can be useful for practitioners. Overall, this thesis demonstrates that (1) such models can be better utilized for unsupervised domain adaptation [1] and (2) conventional wisdom for OOD robustness (see section 3.3) might not apply when amount of fine-tuning data is not as high [2]. We hope to motivate future researchers to also focus on this setting of practical importance.

Appendices

Publications

- **Benchmarking Low-Shot Robustness to Natural Distribution Shifts**

  Aaditya Singh[*,†], Kartik Sarangmath[*1], Prithvijit Chattopadhyay, Judy Hoffman

  Under Review at the 2023 IEEE/CVF International Conference on Computer Vision

- **Adapting Self-Supervised Vision Transformers by Probing Attention-Conditioned Masking Consistency**

  Viraj Prabhu[*], Sriram Yenamandra[*], Aaditya Singh, Judy Hoffman

  2022 Conference on Neural Information Processing Systems (NeurIPS)

[1] $*$ denotes equal technical contribution; † denotes project lead

REFERENCES

[1] V. Prabhu, S. Yenamandra, A. Singh, and J. Hoffman, "Adapting self-supervised vision transformers by probing attention-conditioned masking consistency," *arXiv preprint arXiv:2206.08222*, 2022.

[2] A. Singh, K. Sarangmath, P. Chattopadhyay, and J. Hoffman, *Benchmarking low-shot robustness to natural distribution shifts*, 2023. arXiv: 2304.11263 [cs.CV].

[3] D. Silver *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[4] Q. Rao and J. Frtunikj, "Deep learning for self-driving cars: Chances and challenges," in *Proceedings of the 1st international workshop on software engineering for AI in autonomous systems*, 2018, pp. 35–38.

[5] A. Ari and D. Hanbay, "Deep learning based brain tumor classification and detection system," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 26, no. 5, pp. 2275–2286, 2018.

[6] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning*. The MIT Press, 2009, ISBN: 0262170051.

[7] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR 2011*, IEEE, 2011, pp. 1521–1528.

[8] R. Geirhos *et al.*, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020.

[9] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.

[10] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*, PMLR, 2015, pp. 97–105.

[11] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*, PMLR, 2015, pp. 1180–1189.

[12] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7167–7176.

[13] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," *Advances in neural information processing systems*, vol. 31, 2018.

[14] J. Hoffman *et al.*, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*, Pmlr, 2018, pp. 1989–1998.

[15] D. Kim, K. Wang, S. Sclaroff, and K. Saenko, "A broad study of pre-training for domain generalization and adaptation," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, Springer, 2022, pp. 621–638.

[16] J. Yang, J. Liu, N. Xu, and J. Huang, "Tvt: Transferable vision transformer for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 520–530.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.

[18] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, PMLR, 2020, pp. 1597–1607.

[20] L. Ericsson, H. Gouk, and T. M. Hospedales, "How well do self-supervised models transfer?" In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5414–5423.

[21] M. Caron *et al.*, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[22] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[23] M. Assran *et al.*, "Masked siamese networks for label-efficient learning," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, Springer, 2022, pp. 456–473.

[24] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[25] M. Minderer *et al.*, "Revisiting the calibration of modern neural networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15 682–15 694, 2021.

[26] D. Kim, K. Saito, T.-H. Oh, B. A. Plummer, S. Sclaroff, and K. Saenko, "Cds: Cross-domain self-supervised pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9123–9132.

[27] K. Shen *et al.*, "Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation," in *International Conference on Machine Learning*, PMLR, 2022, pp. 19 847–19 878.

[28] S. Beery, E. Cole, and A. Gjoka, "The iwildcam 2020 competition dataset," *arXiv preprint arXiv:2004.10340*, 2020.

[29] P. Bandi *et al.*, "From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge," *IEEE transactions on medical imaging*, vol. 38, no. 2, pp. 550–560, 2018.

[30] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[31] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," *arXiv preprint arXiv:1705.07204*, 2017.

[32] R. Geirhos, C. R. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann, "Generalisation in humans and deep neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[33] K. Eykholt *et al.*, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1625–1634.

[34] M. A. Alcorn *et al.*, "Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4845–4854.

[35] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *arXiv preprint arXiv:1903.12261*, 2019.

[36] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" In *International conference on machine learning*, PMLR, 2019, pp. 5389–5400.

[37] A. Barbu *et al.*, "Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models," *Advances in neural information processing systems*, vol. 32, 2019.

[38] H. Wang, S. Ge, Z. Lipton, and E. P. Xing, "Learning robust global representations by penalizing local predictive power," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[39] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 262–15 271.

[40] D. Hendrycks *et al.*, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8340–8349.

[41] P. W. Koh *et al.*, "Wilds: A benchmark of in-the-wild distribution shifts," in *International Conference on Machine Learning*, PMLR, 2021, pp. 5637–5664.

[42] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt, "Measuring robustness to natural distribution shifts in image classification," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 583–18 599, 2020.

[43] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.

[44] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," *arXiv preprint arXiv:2202.10054*, 2022.

[45] M. Wortsman *et al.*, "Robust fine-tuning of zero-shot models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7959–7971.

[46] M. Wortsman *et al.*, "Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," in *International Conference on Machine Learning*, PMLR, 2022, pp. 23 965–23 998.

[47] V. Prabhu, S. Khare, D. Kartik, and J. Hoffman, "Sentry: Selective entropy optimization via committee consistency for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8558–8567.

[48] Y. Jin, X. Wang, M. Long, and J. Wang, "Minimum class confusion for versatile domain adaptation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, Springer, 2020, pp. 464–480.

[49] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5018–5027.

[50] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1406–1415.

[51] S. Tan, X. Peng, and K. Saenko, "Class-imbalanced domain adaptation: An empirical odyssey," in *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, Springer, 2020, pp. 585–602.

[52] X. Peng, B. Usman, N. Kaushik, J. Hoffman, D. Wang, and K. Saenko, "Visda: The visual domain adaptation challenge," *arXiv preprint arXiv:1710.06924*, 2017.

[53] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[54] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2414–2423.

[55] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1501–1510.

[56] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," *Advances in neural information processing systems*, vol. 30, 2017.

[57] D. Y. Park and K. H. Lee, "Arbitrary style transfer with style-attentional networks," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5880–5888.

[58] A. Singh, S. Hingane, X. Gong, and Z. Wang, "Safin: Arbitrary style transfer with self-attentive factorized instance normalization," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2021, pp. 1–6.

[59] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," *Advances in neural information processing systems*, vol. 28, 2015.

[60] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, 2017, https://distill.pub/2017/feature-visualization.

[61] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 116–12 128, 2021.

[62] Y. Deng *et al.*, "Stytr2: Image style transfer with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 326–11 336.

[63] F. Phillips and B. Mackintosh, "Wiki art gallery, inc.: A case for critical thinking," *Issues in Accounting Education*, vol. 26, no. 3, pp. 593–608, 2011.

[64] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, Springer, 2014, pp. 740–755.

[65] H. Yang *et al.*, "Domain invariant masked autoencoders for self-supervised learning from multi-domains," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, Springer, 2022, pp. 151–168.

[66] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers amp; distillation through attention," in *International Conference on Machine Learning*, vol. 139, Jul. 2021, pp. 10 347–10 357.

[67] H. Chefer, I. Schwartz, and L. Wolf, "Optimizing relevance maps of vision transformers improves robustness," *arXiv preprint arXiv:2206.01161*, 2022.

[68] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in neural information processing systems*, vol. 33, pp. 9912–9924, 2020.

[69] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[70] Y. Guo *et al.*, "A broader study of cross-domain few-shot learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, Springer, 2020, pp. 124–141.

[71] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola, "Rethinking few-shot image classification: A good embedding is all you need?" In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, Springer, 2020, pp. 266–282.

[72] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[73] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," *arXiv preprint arXiv:1904.04232*, 2019.

[74] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[75] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[76] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 702–703.

[77] Y. Wang, X. Shen, S. X. Hu, Y. Yuan, J. L. Crowley, and D. Vaufreydaz, "Self-supervised transformers for unsupervised object discovery using normalized cut," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 543–14 553.

[78] H. Chefer, S. Gur, and L. Wolf, "Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 397–406.

[79] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "Imagenet-21k pretraining for the masses.[arxiv]," *DOI: https://doi. org/10.48550/arXiv*, vol. 2104, 2021.

[80] M. Assran *et al.*, "The hidden uniform cluster prior in self-supervised learning," *arXiv preprint arXiv:2210.07277*, 2022.

[81] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[82] S. Menon and C. Vondrick, "Visual classification via description from large language models," *arXiv preprint arXiv:2210.07183*, 2022.

[83] T. Brown *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.